Citation and Download Analysis

by

Shaoyi Hu, BComp

A dissertation submitted to

School of Computing

in partial fulfilment of the requirements for the degree of

Bachelor of Computing with Honours

University of Tasmania

November, 2007

Declaration

I declare that this thesis contains no material which has been accepted for the award of any other degree or diploma in any tertiary institution, to my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made in the text of the thesis.

…….

Shaoyi Hu

Abstract

Open access is a mechanism that maximizes the dissertation of scholarly literature world wide, by making them accessible without any financial or other barriers on the Internet. There is a rapid growth in self-archiving articles and publishing them in OA journals, which brings a great impact in research. This research is conducted to investigate whether there is an identifiable causal relationship between downloads and citations of the same articles in an open access repository, as well as the time-varying behaviours of downloading and subsequent citations if possible. To achieve such investigation, an automated monitoring web-based system has been developed for bringing downloading data and citation data together for analysis use. The result suggests that there is no significant pattern found in terms of all the monitored documents in the University of Tasmania open access repository, while this is not always the case if the scope is restricted to only thesis items.

Acknowledgement

First of all, I would like to express my gratitude to all those who gave me the possibility to complete this thesis.

I am deeply indebted to my supervisor, Professor Arthur Sale, for the guidance and suggestions he contributed to this research.

I also wish to thank Lynn Davies, UTas ePrints Libarian, for her advice on improving the accuracy of results.

Thanks go to the School of Computing technical staff: Andrew and Christian for their technical assistance and advice that beyond the call of duty.

Thanks to all my Honours fellows for going through those tough days together.

To all of my housemates, Henry, Chris and Jenny, thanks for your patience and encouragement, as well as Edward looked closely at the final version of the thesis for English style and grammar.

To King, more thanks go to you than I can express. You always put up a lot without complaint. Thanks to Gary for reviewing the draft of thesis with lots of invaluable suggestions, and the food from Timmer and Maggie is always delicious.

Most importantly, this work is dedicated to my parents in appreciation of all the love and support they have given to me. You know I would not be here without you.

Table of Contents

List of Figures

List of Tables

1. Introduction

The failure of library budget in keeping up with the increasing cost of scholarly journals and the advent of the information age have led more and more universities to establish institutional open access repositories. These repositories allow scholarly material to be globally accessible for free via the Internet. Since an open access repository gives a competitive advantage, the research impacts of publications are increased. An OA paper tends to be more likely to be viewed and therefore cited. Existing research indicates that citation of an OA paper leads to 25-250% advantage (depending on discipline and year) than a NOA paper (Hajjem, Harnad & Gingras 2005).These have raised an interesting and unknown question as to whether there is an identifiable causal relationship between the time patterns of downloads from an open access repository, and the subsequent citations.

This thesis will first present a literature review covering topics involved in the research. The discussion on the methodology adopted to test the hypothesis is included in Chapter 2, followed by a chapter that describes the design and implementation of software system developed based on the methodology mentioned in the previous chapter. The results obtained are presented and discussed in Chapter 4, through which the hypothesis of this research is examined. Finally, the conclusions are drawn in Chapter 6 along with the suggestions of how the research might be continued in the future.

2. Literature Review

2.1 Open Access

2.1.1 What is Open Access?

There a variety definitions of Open Access. Budapest Open Access Initiative stated that Open Access is committed to provide free and instant access, for any user world widely, to full-text scholarly literature including unreviewed preprints online (Chan et al. 2002). Arthur Sale (2005) also suggested a similar point: "Briefly, the open access movement is about utilizing the Internet revolution to open the research literature of the world to any user wishing to access it, for free. All that is needed is access to the Internet, and enough bandwidth to download the document". It is important to note that Open Access also requires the subsequent use of publication and is not restricted in subject to proper attribution of authorship (Charles W. Bailey 2005).

2.1.2 Open Access to Scholarly Literature

The open access movement is currently in a state of crisis and is calling for more scholarly literature to be open accessed. This is attributable to the increasing cost of scholarly journals and the failure of library budgets to keep up with these costs. There are two main complementary strategies to achieve open access to scholarly literature as suggested by BOAI (Budapest Open Access Initiative):

- ***Self-Archiving***

  OA Self-Archiving was first introduced by Steven Harnad in 1994 (Poynder 2004). Self-Archiving, referred as to be "Green Road", is a behaviour of depositing a copy of digital document of author's own research paper, including their published articles in NOA journals and unpublished preprints, in their institutional repositories or open archives for the purpose of maximizing its accessibility (Harnad et al. 2004). The right to self-archive preprints is merely a journal policy matter, while the right to self-achieve peer-reviewed article is a legal matter (*Self-Archiving FAQ* 2007).

- ***Open Access Journals***

  The second major strategy, referred to as "Gold Road", are Open Access Journals, in which authors publish their peer-reviewed articles that are accessible without financial or other barriers on the Internet (Harnad et al. 2004). Instead of charging subscription fees as a primarily revenue source for traditional journals, open access journals have various funding strategies. The most common strategies include: "direct author fees, institutional memberships to sponsor all or part of author fees, funding agency payment of author fees, grants to open access publishers, institutional subsidies (such as paying the salaries of journal editorial staff), and priced add-ons (such as recommendation services, current awareness services, or print editions)" (Charles W. Bailey 2005). In addition, the copyright issue is deployed to guarantee open access to all articles published rather than to restrict access to and the subsequently use of the publications (Chan et al. 2002). Testa and McVeigh (2004) proposed that the impact factors of open access journals can be at least as good as those of conventional journals.

A review of the literature up to 2004 reveals that only 5% of journals are gold, while more than 90% of them are green meaning publishers have already endorsed authors' self-archiving of their published peer-reviewed articles. However, only about 10-20% articles have been self-achieved (Harnad et al. 2004).

2.1.3 Open Archives Initiative Protocol for Metadata Harvesting

In order to facilitate the interoperability among different open access archives, institutional repositories and open access journals, the Open Archives Initiative developed a low-barrier mechanism – Open Archives Initiative Protocol for Metadata Harvesting which makes it possible for harvesting records containing metadata from distributed archives or repositories. Therefore, "*Service Provider*" is able to make OAI-PMH service request to retrieve metadata about scholarly work from "*Data Provider*" which maintain repositories having metadata accessible through OAI-PMH interface based on open standard HTTP and XML (*OAI for Beginners: Overview*

2003), to establish an aggregated collection, as illustrated in following figure (Brody, Timothy 2006):

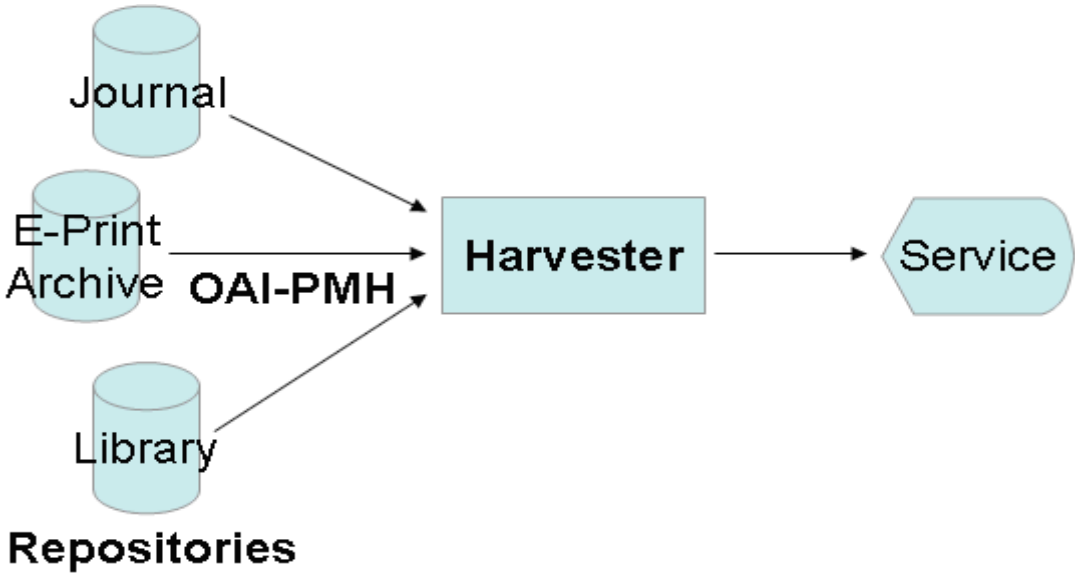


Figure 2-1: High-level OAI-PMH data flow-chart

- ***Dublin Core***

The metadata, encoded in XML, contains a variety of information that describes the deposited documents in open access repositories. The support of the Dublin Core metadata format is essential in OAI-compliant repositories due to the requirement of interoperability. Stuart Weibel asserts that **"**the Dublin Core is a 15-element metadata element set intended to facilitate discovery of electronic resources" (Weibel 1999). While each element can be refined for more specific meaning with qualifiers, only the following 15 unqualified elements are adopted by the OAI-PMH Dublin Core metadata (Brody, Timothy 2006):

Term	Description
contributor	An entity responsible for making contributions to the content of the resource.
coverage	The extent or scope of the content of the resource.
creator	An entity primarily responsible for making the content of the resource.
date	A date associated with an event in the life cycle of the resource.
description	An account of the content of the resource.
format	The physical or digital manifestation of the resource.
identifier	An unambiguous reference to the resource within a given context.
language	A language of the intellectual content of the resource.
publisher	An entity responsible for making the resource available.
relation	A reference to a related resource.
rights	Information about rights held in and over the resource.
source	A reference to a resource from which the present resource is derived.
subject	The topic of the content of the resource.
title	A name given to the resource.
type	The nature or genre of the content of the resource.

Table 2-1: Dublin Core Metadata Element Set

2.2 The Research Impact of Open Access

2.2.1 Correlation of Open Access and Increased Citations

In order to test the impact advantage of open access, Stevan Harnad and Tim Brody (2004) proposed to compare the citation counts of individual OA and non-OA articles appearing in the same (non-OA) journals. It was revealed that there was a dramatic advantage for articles that had been made OA. In their research, the citation impact of 20-40% of articles from the 98% non-OA journals that has been self-archived were compared with all the other articles published in the same journals and the same years in mathematics and physics fields from 1992-2001. The ratios of the two values ranged from 2.5 to 5.8 as increasing by years, which indicated that the ratios would even rise further in the following years.

Rather than be confined with scientific subjects, Hajjem, Harnad & Gingras (2005) used 1,307,038 articles published across 12 years (1992-2003) in 10 disciplines (Biology, Psychology, Sociology, Health, Political Science, Economics, Education, Law, Business, Management) to further test open access the cross-disciplinary

generality of open access impact. They found that the overall percentage of OA articles varied from 5%-16% and was slowly increasing each year. OA articles always attracted more citations, which led to 36%-172% advantage depending on different disciplines and years, compared with NOA articles in the same journals published in the same years. They also estimated the frequencies of citations of OA and NOA articles, which revealed that the annual increase rate of OA articles was considerably higher than NOA articles in every citation range (0, 1, 2-3, 4-7, 8-15, 16+ citations) and this trend was even more significant with those articles having higher citation counts.

2.2.2 Correlation between Citations and Downloads

In order to investigate the relationship between citations and downloads, Brody, Tim, Harnad & Carr (2006) built a "Correlation Generator" which is able to automatically generate statistical graphs for analytical use based on research articles from arXiv.org, illustrated as below:

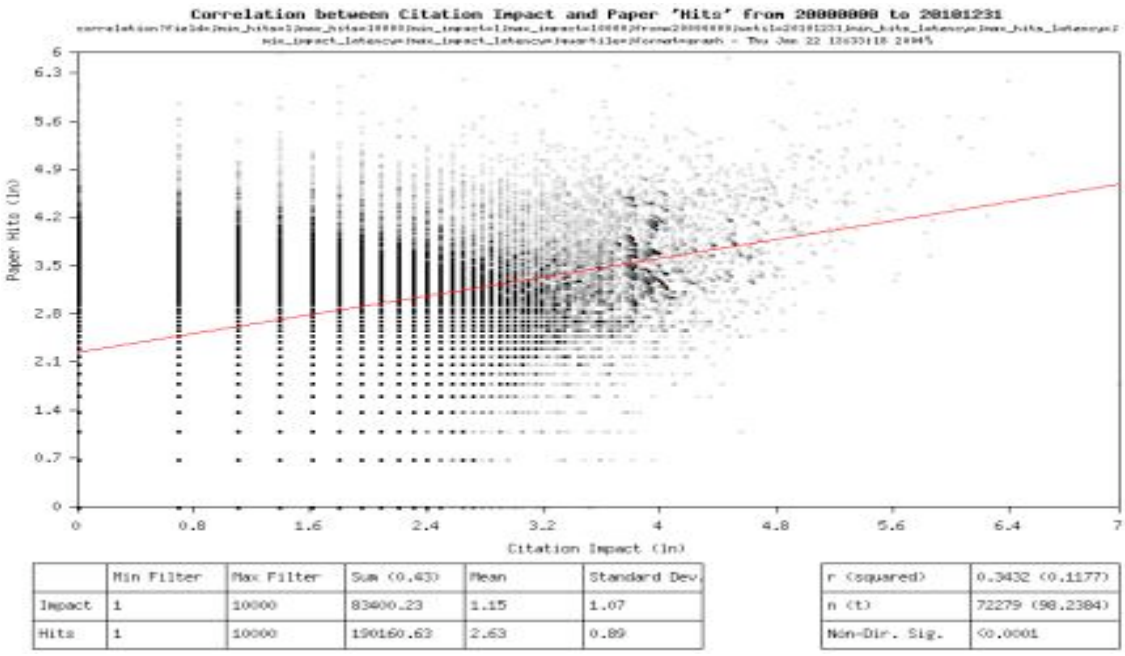


Figure 2-2: Correlation scatter-graph for all papers deposited between 2000-2004

Each dot on the above scatter-graph denotes a monitored article. They proposed that higher citations tend to be associated with higher downloads, even though the overall

frequencies of citations and downloads are scattered. However, this is not always the case depending on different monitored sources.

2.3 The Citation Indexes

2.3.1 ISI Web of Science

ISI web of science is an online academic database provided by Thomson Scientific. It is combined with five databases that include Science Citation Index (SCI), Social Sciences Citation Index (SSCI), Arts & Humanities Citation Index (A&HCI), Index Chemicus, and Current Chemical Reactions. Its databases seamlessly access approximately 8,500 leading articles, including science, technology, social sciences, arts, and humanities (Atkins 1999). The Science Citation Index(SCI) was developed by Eugene Garfield in 1964 (Yancey 2005) and the idea to develop citation index was to "propose a bibliographic system for science literature that can eliminate the uncritical citation of fraudulent, incomplete, or obsolete data by making it possible for the conscientious scholar to be aware of criticisms of earlier papers" (Garfield 1955).

First of all, the Thomson Scientific databases differ from traditional indexing and abstracting services in several ways. The different sorts of citation indexes are multidisciplinary. Consequently, when researchers search for a paper which is indexed by a variety of single-discipline databases, the citation to this paper will include a huge varieties journals in different sorts of subjects (Garfield 1994). Moreover, citation indexes are also comprehensive, not only are originally research papers, review articles, and technical notes provided, but also letters, corrections and retractions, editorials, and other items are provided as well. In addition, "it provides complete coverage of all types of published source items" (Garfield 1994). Last of all, it is apparent that the web of science is only allocated the aggregated link of the papers but not the published page itself. The sample pictures show how the citation index works in Appendix A.

2.3.2 Scopus

Scopus is the largest abstracting and indexing database which is allowed to search for journal articles, books and quality web sources simultaneously (Library 2006). The content of the database covers scientific, technical and medical academic literature organizes from Europe, Latin America and the Asia Pacific region (Raynard 2007). It also contains references and abstracts for articles published in over 15,000 peer-reviewed journals from more than 4,000 international publishers including over 1000 Open Access journals, 500 conference proceedings and over 600 trade publications. The other content Scopus contains shows below:

- "33 million abstracts
- Results from 386 million scientific web pages
- 21 million patent records from 5 patent offices
- Seamless links to full-text articles and other library resources
- Innovative tools that give an at-a-glance overview of search results and refine them to the most relevant hits
- Alerts to keep you up-to-date on new articles matching your search query, or by favorite author" (Elsevier 2007)

The advantage of using Scopus is that it will assist people in getting relevant results expediently and easily. Moreover, Scopus supply tools to sort, redefine and quickly identify results, to make people focus on the outcome of work. In addition, the aim of using Scopus is to spend less time managing the database and more time searching.

2.3.3 Google Scholar

Google Scholar (GS) is a popular freely-accessible web search tool provided by one of the world's largest and most powerful search engine - Google, which developed by Anurag Acharya, an Indian-born computer scientist (Noruzi 2005). It is an incredibly powerful tool which enables searches of scholarly literature, including peer-reviewed papers, theses, books, pre-prints, abstracts, and technical reports. Content of these are from a range of publishers and aggregators with whom Google already has standing arrangements such as universities, academic institutions, professional societies, research groups and preprint repositories (Quint 2004).

Due to the fact that the operation of Google Scholar is based on the Google engine, when publishers and scholarly societies want their content to be accessible via Good Scholar, they just need to contact Google to arrange for Google's spiders to crawl their sites which must provide access for non-subscribers to bibliographic citations and abstracts (Quint 2004). The more easily open access scholarly materials can be accessed by Google Scholar, the lower the price for using academic journals and databases. Therefore, it is beneficial for the researchers to find out the greater valuable scholarly materials. Consequently, Anurag Acharya were correct in their assertion that the goal of Google Scholar was to "make it easier to find content, open access or not. The first step in any research is to find the information you need to learn and then build on that. Not being able to find information hinders scholarly endeavor." (Noruzi 2005).

2.3.4 Comparing among Web of Science, Scopus and Google Scholar

For WoS (Web of Science), the ISI citation databases mainly focus on North American, Western European, and English language titles and the range of search is limited from 8,700 journals which contain in its own database. When the researcher is using WoS, it is unable to list any citations from books and most conference proceedings and occasionally there will be some citation errors in the result of faculty of ISI indexing, such as homonyms, synonyms, and inconsistency in the use of initials and in the spelling of non-English names (I.Meho & Yang 2007).

For Scopus, it can access/reference the largest amount of literature and supplies some human-based services such like sort, redefine, and quickly identify results. All of these services will be beneficial for the preparation of papers. However, suffers from the same issues as the WoS which is only accessible by the specific databases (Elsevier 2007).

In contrast to WoS and Scopus, Google Scholar Google does not offer a publisher list, title list, document type identification, or any information about the time span or the refereed status of records in GS (Bakkalbasi et al. 2006). However, the citation databases covers non-electronic and electronic scholarly materials that make the

searching range of Google Scholar larger compared with others. As such, the key advantage of Google Scholar is to allow researchers to "trace what articles are cited by a particular article and where the article has been cited elsewhere. This can be useful for developing a bibliography or tracing the development of a topic or issue on the Web" (Noruzi 2005) even if Google Scholar sometimes gives duplicated citations.

2.4 Development Tools

2.4.1 PHP

PHP (Group 2007) originally stood for *Personal Home Page* and now stands for *PHP Hypertext Preprocessor*. It is a widely-used open source server-side scripting language designed specifically for the web and can be embedded into HTML. PHP was conceived by Rasmus Lerdorf in 1994 and it is originally designed to run under Linux, using Apache Web server. Until October 2002, it was in use on more than nine million domains worldwide and the number is rapidly increasing. Moreover, PHP can be deployed on most standard web servers and can be used for almost every operating platform such like UNIX, Linux, Windows and Mac OS. The principal reason for the popularity of PHP is that it has these following advantages:

High Performance

PHP is very efficient. Compared with other equivalent languages such as JSP and ASP.NET, PHP pages always take less time to execute. In addition, a single inexpensive server can be used in conjunction with PHP to serve millions of hits per day. Benchmark data published by Zend Technologies (http://www.zend.com) shows PHP is currently outperforming the competition (Welling & Thompson 2003).

Database Integration

PHP can directly connect to Oracle, MySQL, PostgreSQL, dbm and filePro. Furthermore, it can connect to any database which has an *Open Database Connectivity Standard* (ODBC) drive (Welling & Thompson 2003). This means that PHP has "native connections available to many database systems" (Welling & Thompson 2003)

Built-in Libraries

PHP can use many built-in functions to perform many useful web-related tasks. For instance, if there are GIF images or PDF documents which need to be generated by using PHP, only a few lines of code will be required to provide such functionality (Welling & Thompson 2003).

Open Source Code and Portability

The main reason for PHP's popularity is its portability. First of all, PHP is available for many different operating systems. Consequently, the code syntax and the display format are the same on the different systems, no matter whether the system is UNIX or Windows. Furthermore, PHP is an open source language which means that developers can use it without any license and can make modifications without any support. In addition, developers can use PHP every time without worrying about whether "the manufacturer goes out of business and decide to stop supporting the product" (Welling & Thompson 2003).

2.4.2 HTTP Client

HttpClient is a client class for the HTTP (Hyper-Text Transfer Protocol). Firstly it was started in 2001 as a subproject of the Jakarta Commons and out of it in 2003, being an independent top level project in 2007 (*Jakarta Commons HttpClient*). Moreover, HttpClient is used for interacting with another web server from within PHP script. The POST and GET methods are used to interact with a server, as well as retrieving information from a server. It can therefore be used as part of any script that needs to communicate with an application running on another site (Willison 2003). An http client class which Supports (GuinuX 2002):

"HTTP Proxy with Basic Authentication
multipart/form-data AND application/x-www-form-urlencoded
GET, HEAD and POST methods
HTTP cookies
Chunked Transfer-Encoding

HTTP 1.0 and 1.1 protocols

Keep-Alive Connections

Basic WWW-Authentication"

2.4.3 HTML Parser

The HTML Parser is used to parse HTML in either a linear or nested fashion which is used for transformation and extraction by using Java library which include filters, visitors, custom tags and easily use JavaBeans. In addition, "It is a fast, robust and well tested package" (*HTML Parser* 2006).

- **HtmlSQL**

The HtmlSQL is an experimental PHP class which is useful for querying the web by using SQL syntax rather than writing some complex functions. The sample shows below:

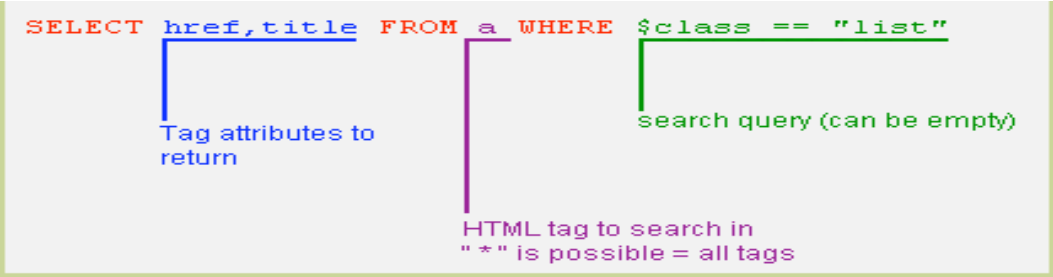


Figure 2-3: An example query of HtmlSQL

This query returns an array with all links that contain the attribute class="list" (John 2006).

2.4.4 JpGraph

JpGraph is an Object-Oriented Graph which is used to create numerous types of graphs, either on-line or written to a file. The advantage of JpGraph is that it makes it easier to draw both "quick" and "dirty" graphs which include a lot of code as well as

complex graphs which require a very fine grained control. “The library is completely written in PHP and ready to be used in any PHP scripts (both CGI/APXS/CLI versions of PHP are supported)” (*What is JpGraph?* 2007). The sample pictures shows below (created by JpGraph Library):

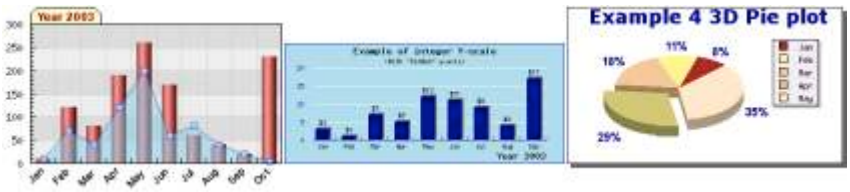


Figure 2-4: Examples of JpGraph

2.4.5 Cron Job

Cron Job is a time-based scheduling service in UNIX which allows tasks to be run automatically in the background in the regular intervals by the cron daemon. These tasks are often termed as cron jobs in unix , solaris and Cron Job driven by a *crontab*, a configuration file that specifies shell commands to run periodically on a given schedule (*Crontab - Quick reference*). There is a sample which includes the explanation to be shown below:

```
*    *  *  *  * command to be executed
-    -  -  -  -
|    |  |  |  |
|    |  |  |  +----- day of week (0 - 6) (Sunday=0)
|    |  |  +------- month (1 - 12)
|    |  +--------- day of month (1 - 31)
|    +------------ hour (0 - 23)
+-------------- min (0 - 59)
```

Figure 2-5: Crontab Specification

Each of the patterns from the first five fields may be either * (an asterisk), meaning all legal values, or a list of elements separated by commas. The subsequent fields (i.e., the rest of the line) specify the command to be run.

3. Methodology

3.1 Introduction

This chapter describes the methodology adopted in testing the hypothesis of this research, as well as addressing the problems of existing system. It also presents an evaluation of whether there is an identifiable causal relationship between the downloads of articles and citation of the same articles from an open access repository, including the investigation of time-varying behaviours of downloading and the subsequent citations if time permits, a web-based system that capable of collecting both of downloading and citation data and producing data updated weekly in automated way is required to be developed. The implementation of the system based on the methodology described is introduced in the next chapter.

UTas eprints is an open access repository that provides free, online, full-text and open access to the research output of the University of Tasmania (*ePrints home*). UTas eprints is also placed in first position based on ratings of university coverage in Australia (Sale 2007), meaning the documents deposited in the repository are from a relatively variety of subjects compared with the others. In addition, the software system is running on an internal server in the School of Computing, so that the data transaction time between the server hosting the software system and the server hosting the repository itself would expedited, which saves time consumption in data collection. Taking the considerations outlined above into account, the UTas eprints is selected as a cornerstone to establish the experiment, from where the relevant data is collect.

3.2 Data Collection Methods

3.2.1 Documents Information

As discussed in Chapter 2, each Open access repository is constructed to provide the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Each document in the repository has its own metadata. Therefore, one way to collect the

information on documents should be monitoring the Metadata accessible through OAI-PMH.

In UTas eprints, there is a variety of information about documents defined in Metadata. Figure 3-1 illustrates the html source of the page from a document that resides in the repository. It can be seen that the document title is defined in the Metadata content having a name attribute identified as “DC.title” and “eprints.title”. It is important to note that these two elements contain the same content, however, the name attributes starting with “eprints.” may change when the version of the repository software gets updated, while those starting with “DC.” (‘Dublin Core’ – a set of international recognised metadata elements for cross-domain information resource description, more details in Chapter 2) stands. Thus, only those “DC” metadata are used when collecting the relevant data from the metadata. The name of the author is defined as “DC.creator”, and the other kinds of information about the document are also defined in the Metadata.

```
<meta content="Sale, Arthur" name="eprints.creators_name" />
<meta content="article" name="eprints.type" />
<meta content="2005-01-04" name="eprints.datestamp" />
<meta content="2007-10-01 02:24:47" name="eprints.lastmod" />
<meta content="show" name="eprints.metadata_visibility" />
<meta content="The Implementation of Case Statements in Pascal"
name="eprints.title" />
<meta content="1981" name="eprints.date" />
<meta content="published" name="eprints.date_type" />
<meta content="Software - Practice and Experience"
name="eprints.publication" />
<meta content="11" name="eprints.volume" />
<meta content="John Wiley &amp;amp; Sons Ltd"
name="eprints.publisher" />
<meta content="929-942" name="eprints.pagerange" />
<meta content="University of Tasmania"
name="eprints.institution" />
<meta content="UNSPECIFIED" name="eprints.thesis_type" />
<meta content="TRUE" name="eprints.refereed" />
<meta content="The Implementation of Case Statements in Pascal"
name="DC.title" />
<meta content="Sale, Arthur" name="DC.creator" />
<meta content="280300 Computer Software" name="DC.subject" />
<meta content="John Wiley &amp;amp; Sons Ltd" name="DC.publisher" />
<meta content="1981" name="DC.date" />
<meta content="Article" name="DC.type" />
<meta content="PeerReviewed" name="DC.type" />
<meta content="application/pdf" name="DC.format" />
<meta content="http://eprints.utas.edu.au/126/1/CaseStmts.pdf"
name="DC.identifier" />
<meta content="Sale, Arthur (1981) The Implementation of Case
Statements in Pascal. Software - Practice and Experience, 11 .
pp. 929-942." name="DC.identifier" />
<meta content="http://eprints.utas.edu.au/126/"
name="DC.relation" />
```

Figure 3-1: The html source page of the page that a document resides

Figure 3-1 also shows that the definition of a document type, defined with a name attribute as “DC.type” , is not only confined to its category, but also indicates whether the document is peer-reviewed or not. This is a very important piece of information. Being consequent on the main objective of this research is to examine the patterns between download and citation, only those kinds of documents which are more likely to be cited are taken into account.

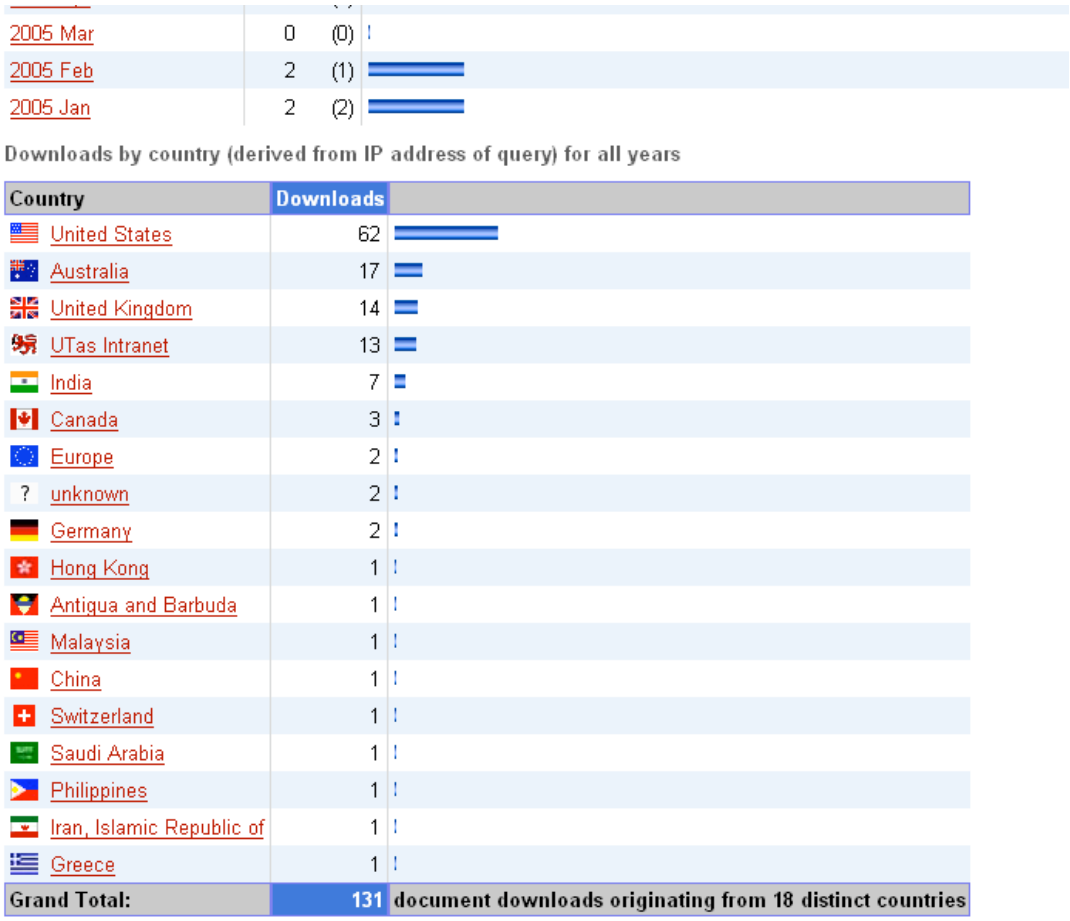
There are several document types defined in UTas eprints, such as “Article”, “Workshop or Conference Item”, “Thesis”, “Book”, “Book Chapter”, “Report” and “Other”, all of which are associated with “PeerReviewd” or “NonPeerReviewd”. When the citation count occurs to “Book Chapter” item, it is hard to differentiate whether the citation is caused by the book of which the chapter is a part or the chapter itself, which confuses the analysis. “Report” and “Other” items, almost of which appear as slides or images, tend to be not cited. The not peer-reviewed “Article” or “Workshop or Conference Item” items are drafts of scientific papers that have not yet been published in peer-reviewed scientific journals or presented in formal conferences, which would not attract citations from other papers. Consequently, only “Article” and “Workshop or Conference Item” are associated with “PeerReviewed”. “Thesis” and “Book” items are also monitored, but not the others.

Furthermore, the ID code of the document from a Metadata defined with a name attribute as “DC.relation” is also necessary to be collected when tracking back the document to the repository. In this stage, the information about its title, author’s name, document type and ID filtered with the defined document types for each document is collected for the later collecting citation data of the document, which is described in the later section.

3.2.2 Downloading Data

The downloading data is also available on a webpage of the repository. There is an outgoing link to a page showing the downloading data varied by times and countries

from where the downloads derive from each page that the document resides, as well as the total downloading data, illustrated in the Figure 3-2 below:



2005 Mar	0	(0)
2005 Feb	2	(1)
2005 Jan	2	(2)

Downloads by country (derived from IP address of query) for all years

Country	Downloads	
United States	62	
Australia	17	
United Kingdom	14	
UTas Intranet	13	
India	7	
Canada	3	
Europe	2	
unknown	2	
Germany	2	
Hong Kong	1	
Antigua and Barbuda	1	
Malaysia	1	
China	1	
Switzerland	1	
Saudi Arabia	1	
Philippines	1	
Iran, Islamic Republic of	1	
Greece	1	
Grand Total:	**131**	**document downloads originating from 18 distinct countries**

Figure 3-2: The total downloading data in the webpage

Therefore, the downloading data page for a particular document is identified in respect with its ID code defined in the last section. For example, if the ID code of the document is defined to be 126 (refer to the URL of document webpage - http://eprints.utas.edu.au/126/), the URL of the downloading data webpage should be http://eprints.utas.edu.au/es/index.php?action=show_detail_eprint;id=126;, which contains the same ID code accordingly and associated with the other parts of the URL that always stand over different downloading data pages.

The total downloading data, then, is collected through the html source page of the downloading data page containing the consistent downloading data with it is in the webpage as shown in Figure 3-3 in this stage.

```
style="font-size:small;border-right:1px solid #dddddd;">1</td><td
align="left"><img src="bars/hp.png" alt="views" height="9"
width="1px"></td></tr><tr><th rowspan="1" colspan="2"
style="background-color:#cccccc;">Grand Total:</th>
                <th align="right" style="background-
color:#4477dd;color:#ffffff;">131</th><th align="left"
```

Figure 3-3: The total downloading data in the html source page

3.2.3 Citation Data

With the competitive advantage of Google Scholar in searching scholarly literature available on the Web and providing the citation index to them as discussed in Chapter 2, it is utilized as the source of citation data for the documents from the repository in this research.

There are a variety of ways to search documents through Google Scholar. In order to search by the title of document, the double quotation mark is placed around the title. As a result, the result page that returned by Google Scholar contains the document (if any) and other documents that mention the title (*Google Scholar Help* 2007), making it difficult for the software to locate the citation data, illustrated as below:

R Milner, J Parrow, D Walker - Information and Computation, 1992 - portal.acm.org
... Mark Hepburn , David Wright, **Trust in the pi-calculus**, Proceedings of the 3rd ACM SIGPLAN international conference on Principles and practice of declarative ...
Cited by 798 - Related Articles - Web Search

[BOOK] Trust in the lambda-Calculus
J Palsberg, P Ørbæk - 1995 - Springer-Verlag London, UK
Cited by 45 - Related Articles - Web Search - BL Direct

From pi-Calculus to Higher-Order pi-Calculus-and Back - all 2 versions »
D Sangiorgi - Proceedings of the International Joint Conference CAAP/FASE ..., 1993 - portal.acm.org
... Mark Hepburn , David Wright, **Trust in the pi-calculus**, Proceedings of the 3rd ACM SIGPLAN international conference on Principles and practice of declarative ...
Cited by 40 - Related Articles - Web Search - BL Direct

A simple view of type-secure information flow in the/spl pi/-calculus - all 12 versions »
F Pottier - Computer Security Foundations Workshop, 2002. Proceedings. ..., 2002 - ieeexplore.ieee.org
Page 1. A Simple View of Type-Secure Information Flow in the -Calculus François Pottier INRIA E-mail: Francois.Pottier@inria.fr Abstract ...
Cited by 34 - Related Articles - Web Search

[PDF] A privacy analysis for the π-calculus: The denotational approach - all 3 versions »
B Aziz, GW Hamilton - Proceedings of the 2 ndWorkshop on the Specification, ... - doc.ic.ac.uk
Page 1. A Privacy Analysis for the π-calculus: The Denotational Approach B. Aziz and GW Hamilton School of Computer Applications ...
Cited by 9 - Related Articles - View as HTML - Web Search

Trust in the pi-calculus - all 4 versions »
M Hepburn, D Wright - Proceedings of the 3rd ACM SIGPLAN international conference ..., 2001 - portal.acm.org
Page 1. **Trust in the Pi-Calculus** Mark Hepburn mark h@postoffice.utas.edu. au David Wright David.Wright@utas.edu.au School of Computing ...
Cited by 6 - Related Articles - Web Search

Figure 3-4: The result page by searching title

In order to further restrict the exact phrase of the query term so that it only appears in the title but not anywhere else, the "allintitle:" operator should be placed in front of the double quoted title. This guarantees Google Scholar will only return results that include the search term in the document's title (*Advanced Scholar Search Tips* 2007). However, for those documents with short titles, the results returned are often erratic, as shown below:

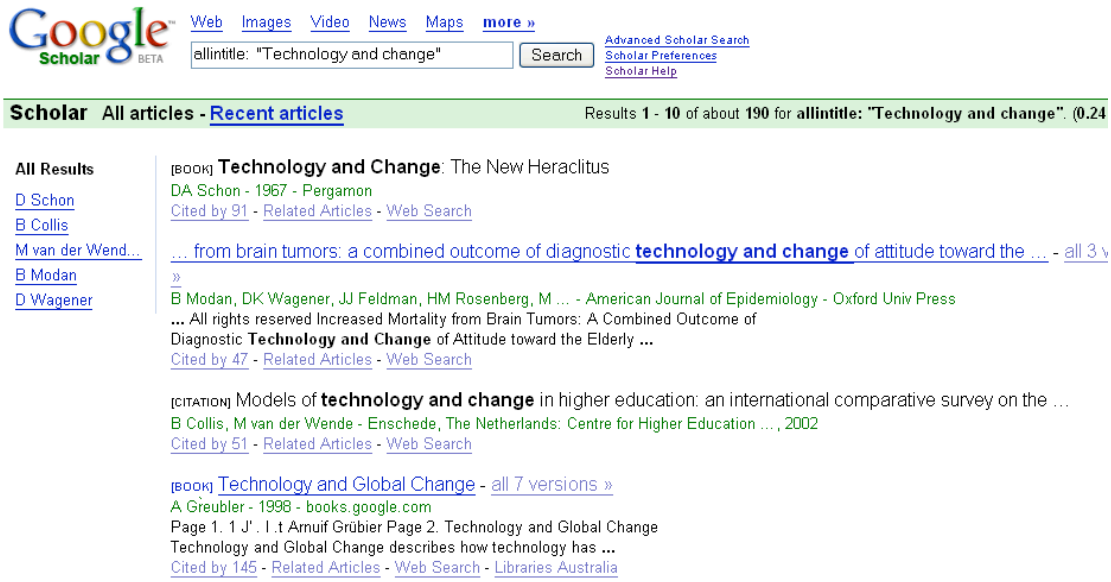


Figure 3-5: The result Page by searching exact title

Therefore, the combination of title and author's name of the document is necessarily to cross define the document, which restricts the document in the returned result not only to matching the title, but also to being written by the author provided in the search query.

To do this, the title and the name of the first author of a multi-authored document defined in section 3.2.1 are passed to Google Scholar's Advanced Scholar Search to locate the document itself in Google Scholar, then obtaining the citation data as provided by "cited by" feature. An example illustrates the procedures of the use of Google Scholar's Advanced Scholar Search shown in Figure 3-6 as below:

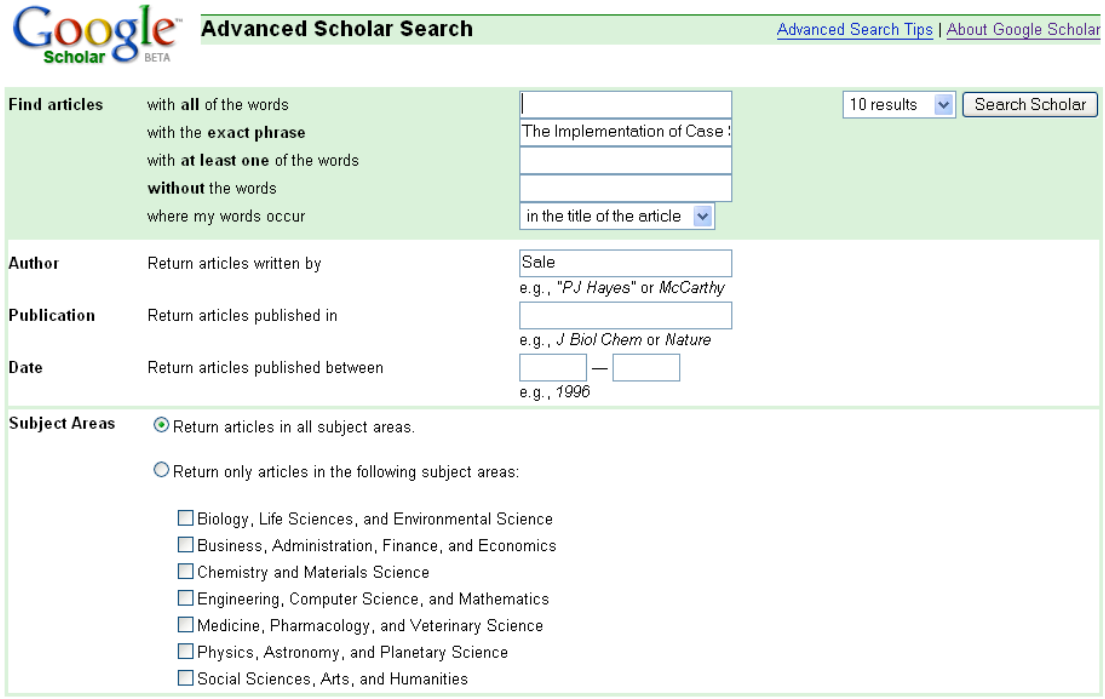


Figure 3-6: Google Scholar's Advanced Scholar Search

It is important to note that only the surname of the name of the author is adopted to fill in the form in Figure 3-6. Noruzi(2005) argues that results are often erratic when the search is made by the entering the author's full name, since the first name appears as an initial in some cases. Only using surname makes it possible to find all authors with that surname.

Then Google Scholar turns out the result page showing the search document as the first item in results if it is found, as illustrated in Figure 3-7 below:

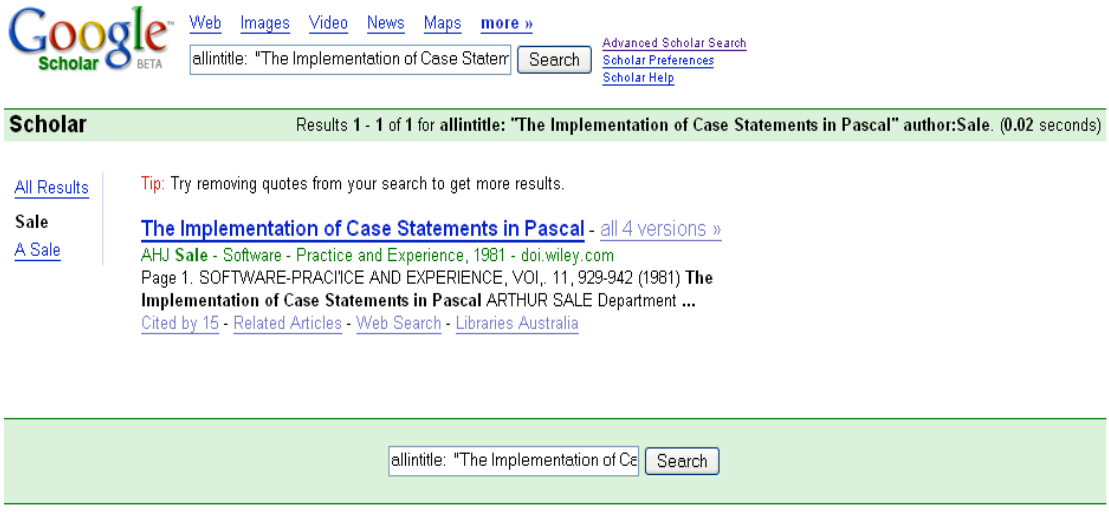


Figure 3-7: The result page of Google Scholar

The number of the times that the document cited, then, is seen from a “Cited by” link in that page and extracted from its html source page (Figure 3-8).

```
<b>The Implementation of Case Statements in Pascal</b></a></span> -
<a class=fl
href="/scholar?hl=en&lr=&newwindow=1&safe=off&cluster=271682295246233
8460" target=nw>all 4 versions &raquo;</a><font size=-1><br><span
class="a">AHJ <b>Sale</b> - Software - Practice and Experience, 1981
- doi.wiley.com</span><br>Page 1. SOFTWARE-PRACI&#39;ICE AND
EXPERIENCE, VOI,. 11, 929-942 (1981) <b>The</b><br>
<b>Implementation</b> <b>of</b> <b>Case</b> <b>Statements</b>
<b>in</b> <b>Pascal</b> ARTHUR SALE Department <b>...</b>
<br><a class=fl
href="/scholar?hl=en&lr=&newwindow=1&safe=off&cites=27168229524623384
60" target=nw>Cited by 15</a>
```

Figure 3-8: The html source page of the result page of Google Scholar

Google Scholar otherwise returns a result page indicating no document is found based on the entered query, as illustrated below:

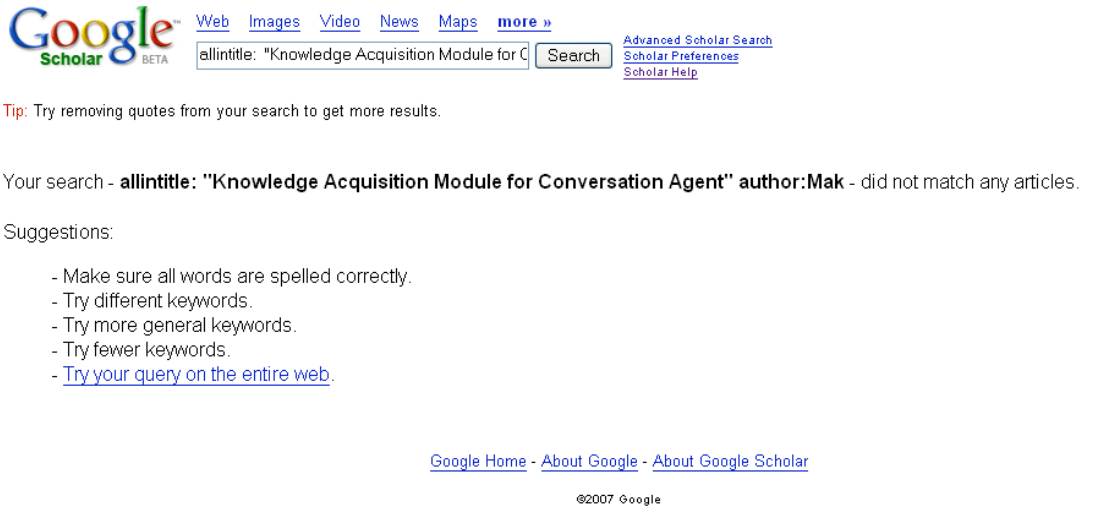


Figure 3-9: The result page indicating no found document in Google Scholar

3.2.4 Time-varying Data

Due to the investigation of time-varying behaviours of downloads and the subsequent citations, in other words, the decay of downloads and citations is a part of the hypothesis of this research, the automatic production of data updated at regular intervals is done by establishing a Cron Job (See Chapter 2). The program is set up to run every week to collect the accumulated data in this research as the time limitation.

4. Software Design and Implementation

This chapter describes the design and implementation of software system based on the methodology that discussed in the last chapter. The detailed description of each step is explained in the following sub sections:

4.1 Overview of Architecture

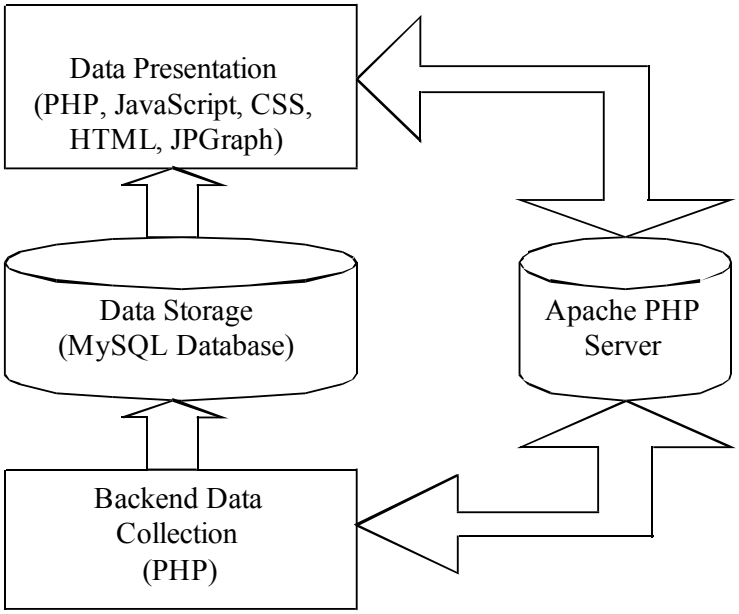


Figure 4-1: The architecture of the software system

As discussed in the last chapter, the document information was defined in the Metadata accessible through OAI-PMH interface, while the downloading data and citation data existed on the relevant web pages themselves. In order to extract downloading data for documents from the web pages of the repository and their citation data from the web pages of the Google Scholar, then bring both of them together for analysis use, an automated software system that implements such function is developed by PHP script running on Apache Server, with integration of a MySQL database.

The software system is designed as a three-layer architecture shown in Figure 4-1. It includes Data Collection Layer (bottom layer) which performs all the relevant data collection in the backend, and then stores those data into a database in its Data Storage Layer (middle layer). With the works done at these two layers, the data is

retrieved from the database and presented for the analysis use at the Data Presentation layer (top layer). The detailed description of each layer is given in the following sections.

4.2 Data Collection Layer

As mentioned in section 3.2, the document information is extracted from the metadata that lies in the html source page of the web pages that describe the document in the repository. Subsequently that information is passed to Google Scholar to be searched to obtain its citation data and used to correspond to the relevant downloading data page in the repository to identify the downloading data. Therefore, extracting the relevant data from web pages is the first phase of the software implementation.

At this layer, the implemented system needs not only to monitor both the repository and Google Scholar, but also to parse the html source code of the relevant web pages to extract the relevant data.

- ***HTTP Client***

  In order to monitor the relevant websites, communication must be established with the servers that host those websites. In respect of UTas eprints and Google Scholar, both are constructed to support the HTTP protocol, an Advanced HTTP Client (See Chapter 2), which is an open source PHP class developed by GuinuX (2002), and is adopted to interact with those servers from within a PHP script in this software system.

  The connections with the UTas eprint server and Google Scholar server are set up by giving URLs within the website. It behaves as a web browser, which automates the process of retrieving web page content from the repository by sending HTTP requests (as many as necessary), then posting forms to Google Scholar to get the result pages returned via GET or POST. The class is also able to distinguish web pages having error code in the head of them, so that those web pages can simply be ignored to save processing time. In addition,

since the software is running on the server hosted by the School of Computing, the proxy authentication that is supported by this class needs to be set up.

- ***HTML Parser***

Once the web page content has been retrieved by HTTP Client, we need a HTML Parser to extract the relevant data out from the HTML source code of that web page. HtmlSQL, an experimental PHP class that query websites or HTML code with an SQL-like Language created by Jonas John (See Chapter 2), is utilized as a part of scripts to implement such function. It makes the software be capable of extracting the data between a pair of defined html tags (<tag>…</tag>).

According the discussion in section 3.2, the first concern is taken into the extraction of document information. The following figure shows how to make HtmlSQL to extract the title of the document, the document type, the document ID code and its authors in the metadata from the webpage that the document resides in the repository.

```
'SELECT content FROM meta WHERE $name=="DC.title" or
$name=="DC.type" or ($name=="DC.relation" and
preg_match("/^http: \/\/eprints.utas.edu.au \/[0-9]+\/$/i", $content)) or
$name=="DC.creator" '
```

Figure 4-2: An example of extracting document information

When document ID code has been extracted in the last phase, it is used to correspondingly locate the relevant downloading data page in the repository (http://eprints.utas.edu.au/es/index.php?action=show_detail_eprint;id=*document ID*). Then, extracting the downloading data of the document from that page would step into the second phase.

```
'SELECT text FROM th WHERE preg_match("/^[0   -9]+$/", $text)'
```

Figure 4-3: An example of extracting downloading data

The earlier collected title and author's name of document is also passed to Google Scholar to be searched to get the result page returned. The final phase is to extract citation data of the document from the result. As Figure 3-7 illustrated, the result page indicating no found document has a hyperlink - Try your query on the entire web, base on which the software is able to check whether there is a result page returned by Google Scholar:

```
'SELECT text FROM a WHERE preg_match("/^Try your query on the entire
web/i", $text)'
```

Figure 4-4: An example of checking no found document

If no result returned by Google Scholar, the citation count is defined as "*NULL*". The following figure shows how the HTML Parser extracts the citation data if the target document is found in the return page.

```
'SELECT text FROM a WHERE $class=="fl" and preg_match("/^Cited by
[0-9]+$/i", $text)'
```

Figure 4-5: An example of extracting citation data

In case of no citation occurred to the target document, the citation count is defined as 0.

In order to implement the data collection task, an external php file needs to be created that have access to both HTTP Client and HTML Parser to obtain the record, including document ID, document title, document author's name, document's downloading data and document's citation data, for each document. It is important to note that, as mentioned in section 3.2.1, only peer-reviewed articles and conference items, thesis and book items are useful for the analysis. Therefore, the downloading

data and citation data for those of documents which types are not one of above defined types would not be tracked, and therefore not be stored into the database.

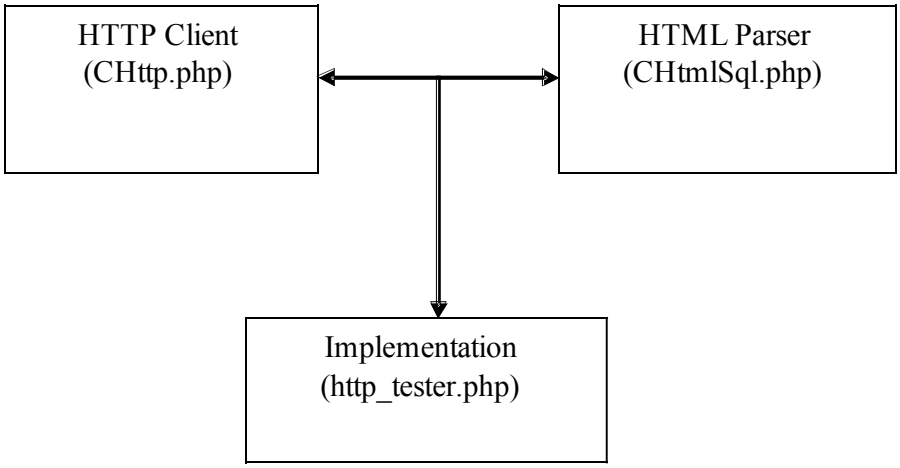


Figure 4-6: The implementation of HTTP Client and HTML Parser

The following pseudo code shows the procedure of implementing HTTP Client and HTML Parser to collect the relevant data and insert into database then.

Begin

1. Call HTTP Client to setup the connection with the repository
2. Assign the range of documents to be monitored
3. For each document, starting from the first one

    if the document exist then

        call the HTML Parser to extract the document info

        if the document type is valid then

        - pass the document ID code to locate the downloading data page accordingly, and extract the downloading data
        - call HTTP Client to setup the connection with Google Scholar
        - pass the document title and the surname of the first author to Google Scholar to get the result page return

            if no document found in Google Scholar then

                the citation count is defined as "NULL"

            else

                if no citation data is found then

                    the citation count is defined as 0

                else

                    extract the citation data

        - insert the relevant data into the database

    elseif the document no found (based on the http error code, eg., 404) then

        ignore the current page and go on to the next page

4. Repeat the above processes over the range defined until no more document exist

End

4.3.1 Database Design

In this software system, all the records are stored in a single table – "*repository_records*" which includes the following information:

- id – The record ID of each document in the database
- doc_id – The document ID code
- title – The document title
- author – The author's name of each document
- download – The downloading data of each document
- citation – The citation data of each document
- doc_type – The document type
- data – The monitoring time

The Table 4-1 shows the table structure of "*repository_records*":

Field	Type	Null	Default	Value and Meaning
id	int(16)	No		auto_increment
doc_id	int(16)	No		
title	text	No		
author	varchar(32)	No		
download	int(16)	Yes	*NULL*	
citation	int(16)	Yes	*NULL*	NULL: Not Found
doc_type	int(16)	No		0: Article 1: Conference Paper 2: Thesis 3: Book
date	date	No		

Table 4-1: The table structure of "repository_records"

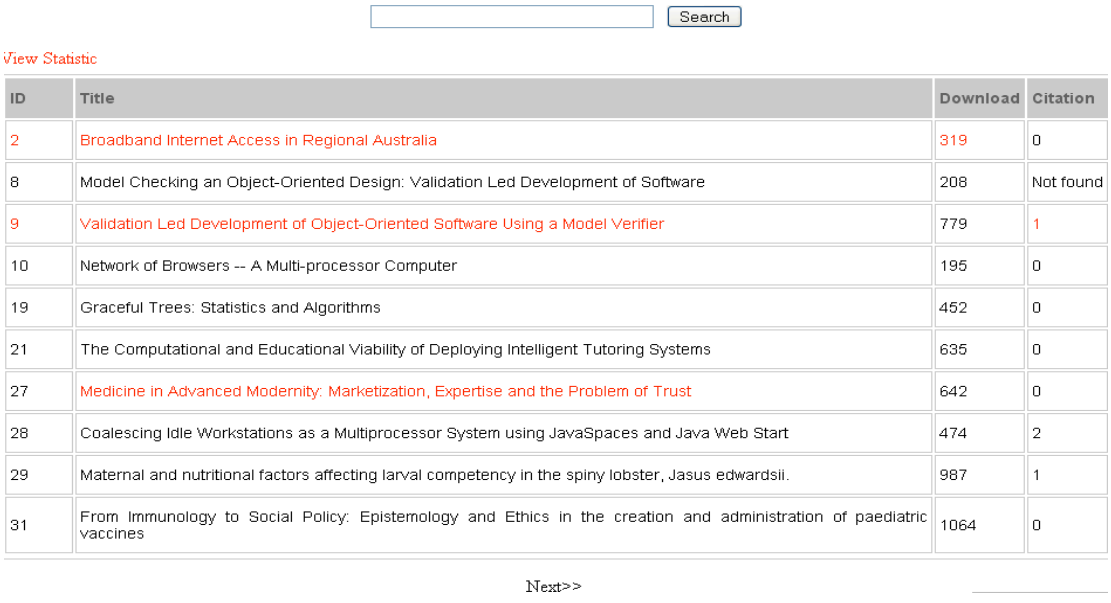
It would be desirable to have document deposit time to be stored into the database, so that the latency or decay of download or citation of each document since they have been deposited could be tracked. Furthermore, with the availability of information on the subjects to which documents belong and the publications in which documents appear, the analysis of patterns of citations and downloads varied by different aggregated categories would become possible. In addition, an examination of the differences between earlier published articles and recent ones requires the publication time to be collected. However, all of these will be conducted in the further work due to the time limitation issue.

4.4 Data Presentation Layer

The front layer is the Data Presentation Layer in which the data is retrieved from the database and presented for the analysis use. It consists of three main components, which are *"Records of Documents View"*, *"History Records View"* and *"Statistics View"*. The detailed description of each component is explained as following:

4.4.1 Records of Documents View

It shows the last updated results of both downloading data and citation data of all the monitored documents, as illustrated in a very small sample in Figure 4-7:


Search

View Statistic

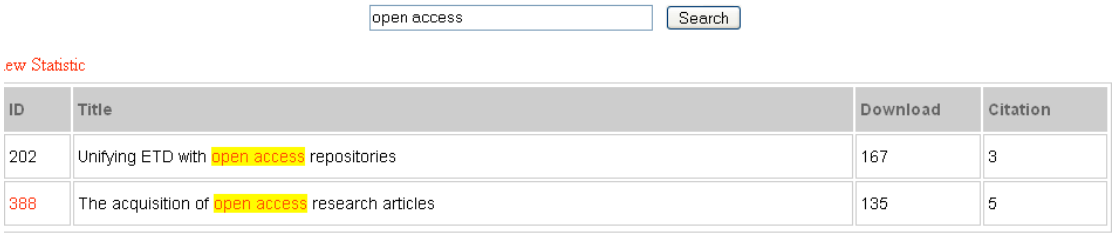
ID	Title	Download	Citation
2	Broadband Internet Access in Regional Australia	319	0
8	Model Checking an Object-Oriented Design: Validation Led Development of Software	208	Not found
9	Validation Led Development of Object-Oriented Software Using a Model Verifier	779	1
10	Network of Browsers -- A Multi-processor Computer	195	0
19	Graceful Trees: Statistics and Algorithms	452	0
21	The Computational and Educational Viability of Deploying Intelligent Tutoring Systems	635	0
27	Medicine in Advanced Modernity: Marketization, Expertise and the Problem of Trust	642	0
28	Coalescing Idle Workstations as a Multiprocessor System using JavaSpaces and Java Web Start	474	2
29	Maternal and nutritional factors affecting larval competency in the spiny lobster, Jasus edwardsii.	987	1
31	From Immunology to Social Policy: Epistemology and Ethics in the creation and administration of paediatric vaccines	1064	0

Next>>

Figure 4-7: Records of Documents View

The records of the rest of monitored documents can be viewed via the paging navigation – "Next>>" at the bottom of the page. Each *"Download"* and *"Citation"* cell is clickable which redirect users to the original web pages from where the data was collected. By clicking the link in the *"ID"* cell, the web pages that the document resides in the repository is tracked. An important feature provided in this page is that the page showing the history records of each document is able to be redirected by clicking the link in the *"Title"* cell. Furthermore, there is a search feature that allows

users to search a particular document by entering a set of keywords at the top of the page.



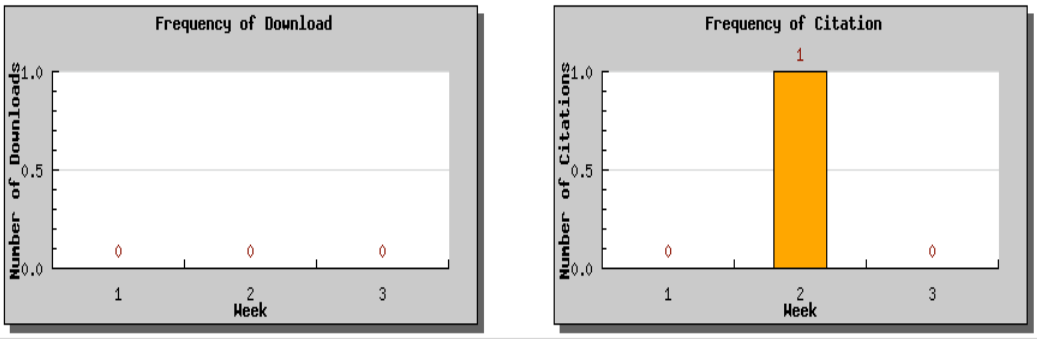
ID	Title	Download	Citation
202	Unifying ETD with open access repositories	167	3
388	The acquisition of open access research articles	135	5

Figure 4-8: The search feature

As shown in Figure 4-8, the documents with titles or parts of them matching the search query are returned in the result page.

4.4.2 History Records View

From the history records page, all the history records for a particular document are able to be traced back according to each monitoring time. The following figure shows an example of document having data changed during the monitoring period:



ID	Title	Download	Citation	Update Date
1911	Dual Symmetric Inverse Monoids and Representation Theory	0	13	2007-10-29
1911	Dual Symmetric Inverse Monoids and Representation Theory	0	13	2007-10-22
1911	Dual Symmetric Inverse Monoids and Representation Theory	0	12	2007-10-15
1911	Dual Symmetric Inverse Monoids and Representation Theory	0	12	2007-10-8

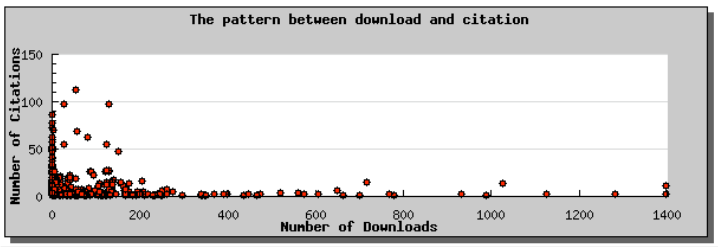
Figure 4-9: History Record View

The two graphs, generated dynamically depending on the change of the data by using JPGraph (See Chapter 2), illustrate the frequency of download and citation varied by each monitoring time, as a result, the time-varying behaviour can be examined by analysing these graphs.

4.4.3 Statistics View

Since the main objective of this research is to provide data for the analysis use, some of statistics tables and graphs are required to be generated dynamically with the last updated data. In the statistics page, the patterns between download and citation and the distribution of download and citation are able to be provided to users to analyse, as illustrated in a very small sample below:

Total



Total of Documents	Download	Citation
939	96524	5036

Download

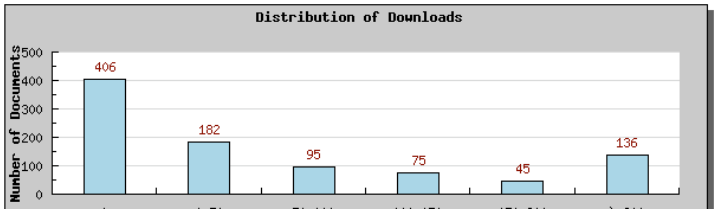


Figure 4-10: Statistics View

5. Results and Discussion

The following chapter examines some results that can be derived from the data produced by the software. Many different analyses are possible, to determine the patterns between downloads and citations, the distribution of downloads and citations, and to analyse the contents of the repositories. The key issue is the bringing together of two pieces of public information about a document; the analyses follow from the merging of this information.

Since the length of the Honours program did not permit collecting a long time series of data, the analyses that can be performed with the current data are limited. Only four weeks of data has been accumulated which is far too short to analyse any significant citation trends, which generally develop over periods of years.

5.1 The Patterns between Downloads and Citations

The scatter graph below, generated by the software dynamically, is a snapshot of the patterns between downloads and citations, as well as the basic distribution of them, for *all* the monitored documents in the repository. The downloading and citation data are the accumulated amounts up to the 29th of October 2007.

In the scatter graph, the X-axis denotes the number of downloads, whereas the citations data lies on the Y-axis. The scatter graph consists of dots, each of which represents a monitored document with its downloading and citation data accordingly. Those documents having zero citation or 'not found' citations are excluded. From the distribution in the graph, it can be seen that most of dots lie near the origin, which reflects the majority of documents are neither highly downloaded nor highly cited. The low citation level is typical of papers in general (often 80% of papers are never cited), while the zero downloads may derive from either the recency of the upload (the repository has been very active in the latter part of 2007 due to the RQF), or may be due to the fact that there is no full-text to be downloaded. In a later section, the confusing factors in the UTas repository are discussed.

In general, the documents with higher value of downloads do not tend to have higher citation value, while some of documents have high numbers of citations but only a few downloads. It is important to note that there are a few of dots that lie on the Y-axis, in other words, those documents are cited, but not downloaded. Those documents might have been published in other print publications before they have been deposited in the repository, and might have been cited by other papers from viewing their original publications.

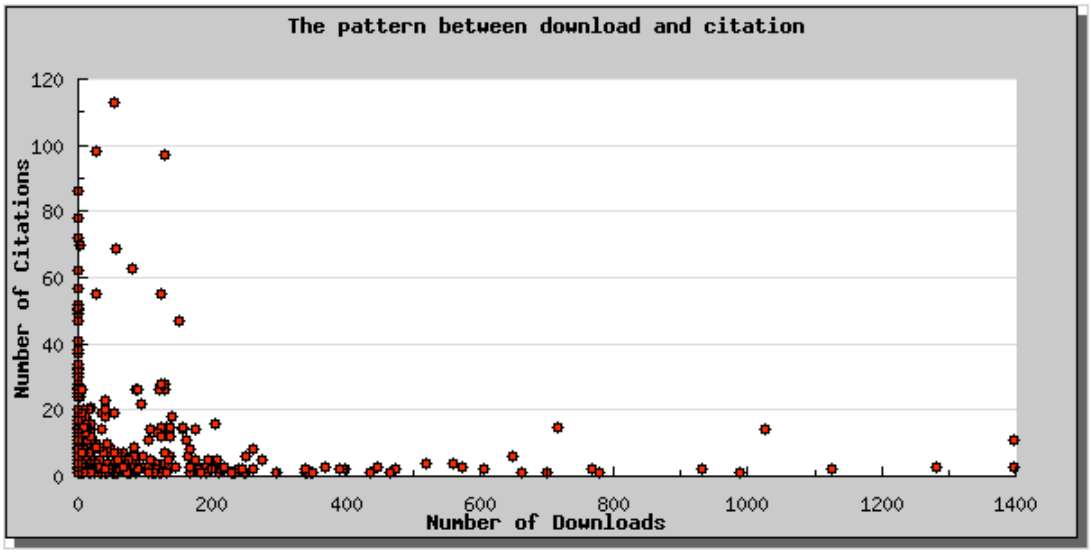


Figure 5-1: The pattern between download and citation

5.2 The Distribution of Downloads

The following bar graph illustrates the distribution of downloads, which is dynamically generated with the last updated data. The Y-axis denotes the number of documents, while the range of each block of downloading data is indicated on the X-axis. Each column corresponds to the number of documents having downloading data within the range of each block. It can be seen that the number of documents with zero downloading data are many more than those lie in the other blocks. See the later section for the analysis of this situation. For non-zero downloads, the number of documents in the category decreases as the number of downloads increases, and this is typical of research articles. It is also related to the length of time the article has been uploaded to the repository. It can also be seen from Table 5-1 that the distribution of

downloads is very broadly scattered according to the value of coefficient variation. However, the distribution is not Gaussian, and these measures are only roughly applicable.

mean	102.7945
standard deviation	232.3306
coefficient variation	226.01%

Table 5-1: Standard deviation of downloads

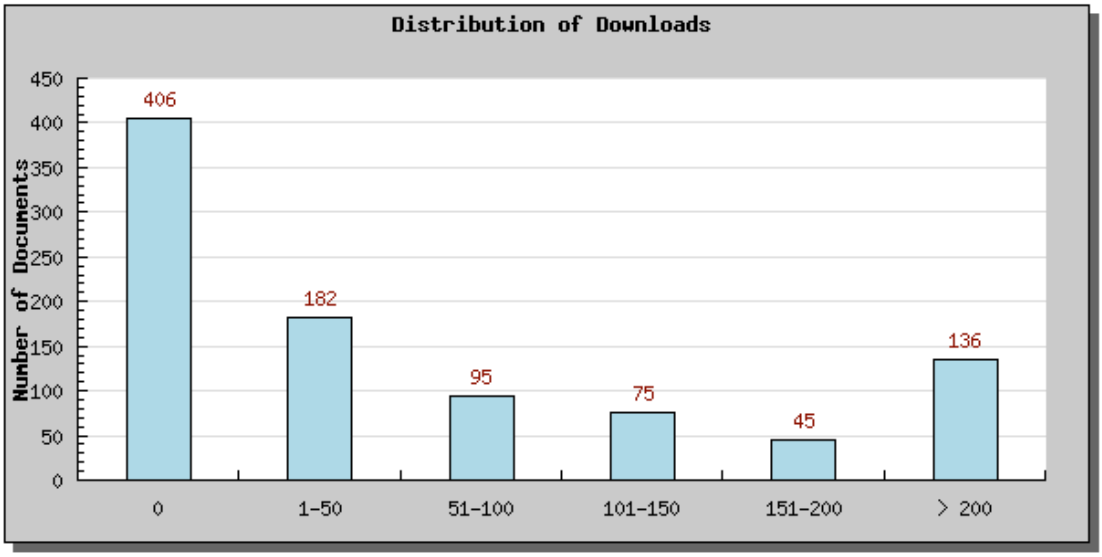


Figure 5-2: Distribution of Downloads

5.3 The Distribution of Citations

With the last updated data, the bar graph indicates the distribution of citations is shown in the following figure. The Y-axis denotes the number of documents, while the range of each block of citation data is indicated on the X-axis. Each column corresponds to the number of documents having citation data within the range of each block. The first block indicates those documents that could not be found in Google Scholar. It can be seen that the most of documents either have no citation number or a small citation number. This is typical of research articles. The number of documents

drops significantly when the citation number is more than 5, after then the number of documents keeps going down slowly as the number of citation increases. In addition, the Table 5-2 reflects the fact that the distribution of citation is also very large according to the value of coefficient variation. Again, the distribution is not Gaussian, and the coefficient of variation is only crudely applicable.

mean	6.263682
standard deviation	12.11804
coefficient variation	193.47%

Table 5-2: Standard deviation of citations

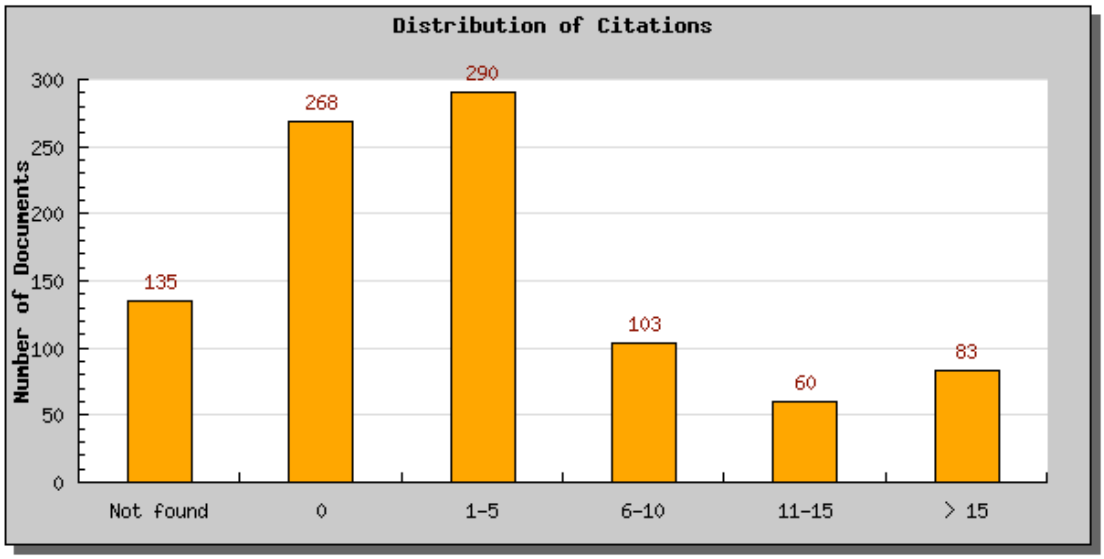


Figure 5-3: Distribution of Citations

5.4 Demonstration of Versatility

The data are capable of being analysed in many other ways. As previously mentioned, the articles in the UTas repository have many characteristics that confuse simple analyses of the contents, For example:

1. There are a significant number of articles whose full-text is *Restricted* (in other words not available for download). The download data for all these articles is naturally always zero.
2. Many articles are published in subscription (print) journals and then deposited into the repository. If they have acquired citations, this may be because the citing authors saw the article in the paper journal, rather than online.
3. In some cases, authors place old articles in the repository for completeness of their record – much the same as the previous case but the deposit is much delayed after publication, perhaps as much as 20 years. The citation count for such items has little to do with the open access repository and its downloads. However, the downloads can be viewed as an expression of continuing interest in the item.
4. There are disciplinary differences in both citation and download behaviour. Some disciplines cite more; and the general public (as well as researchers) are more interested in some disciplines than others.

To demonstrate the versatility of the data, it was also decided to re-analyse the data only for PhD, Master and First Class Honours theses deposited in the repository. Theses are seldom published in their entirety in print, and all the theses in the repository are relatively recent. Few (if any) theses are Restricted. Thus confusing factors 1-3 disappear. Virtually the only way a thesis can be cited, unless by its author, is through open access.

The results of this analysis for theses are shown below. A similar scatter graph is presented.

- **Patterns of Thesis**

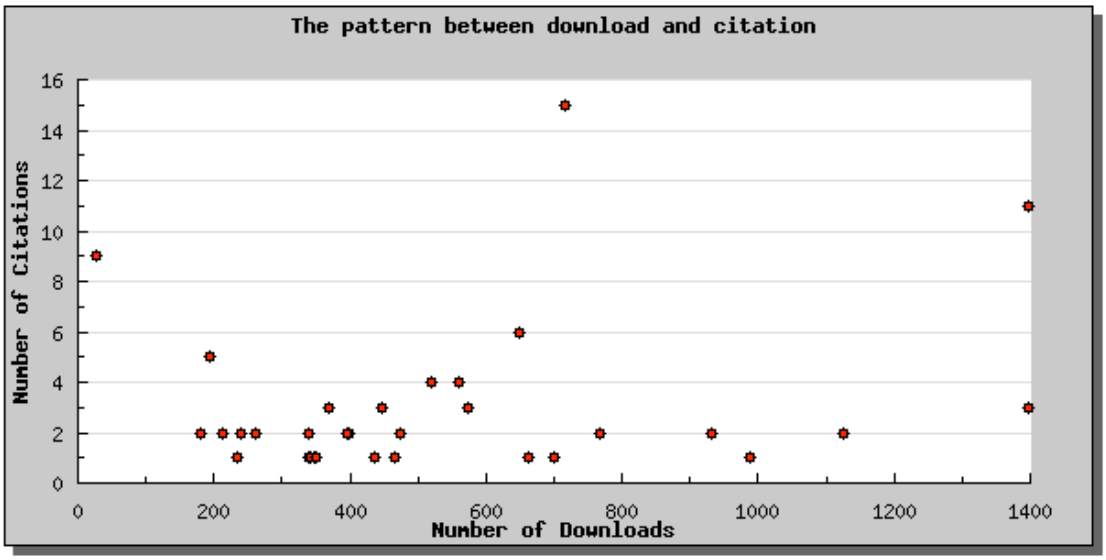


Figure 5-4: The pattern between download and citation of thesis

It can be seen that every thesis in this collection has at least one citation. There is a general trend for higher citations to be associated with higher download frequency; for example above 1000 downloads no thesis has a single citation. It would be desirable to do more work on analysing this data by discipline, and length of time in the repository. Identifications of self-citations (ie citations by the thesis author) could be identified. However, there was insufficient time for this further work.

For interest the same graphs were produced for (a) refereed journal articles, (b) refereed conference papers and (c) books. The diagrams show the expected patterns. Citations of journal articles appear to be dominated by prior publication. Two journal articles appear to be very heavily downloaded, and further work might identify why. Conference papers are not widely disseminated in print, and the graph shows a correlation between downloads and citations. It is apparent that citation levels for conference papers are generally low. This may be because most of the conference papers were in computer science, or because in other disciplines conference papers are often not cited. Furthermore, since there is only one book that shows up in the graph, the patterns of book cannot be examined.

- **Patterns of Articles**

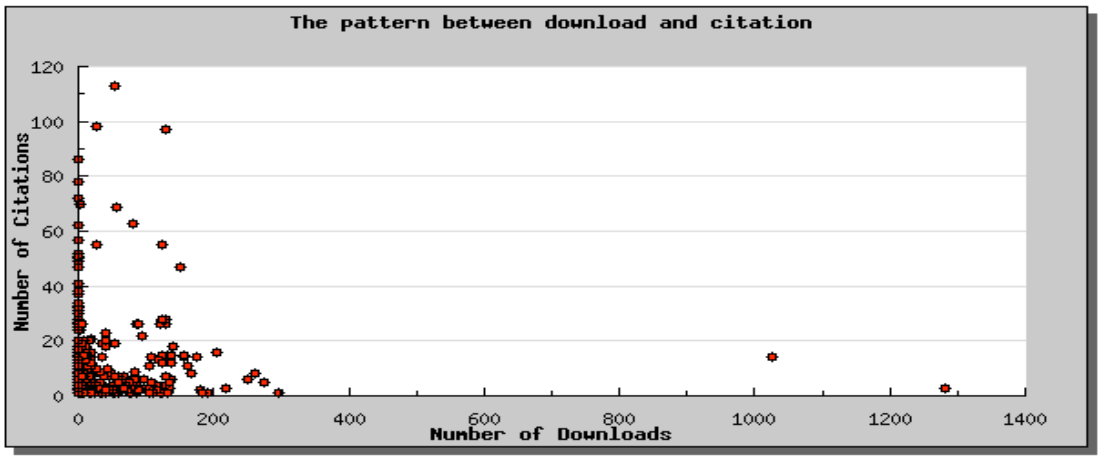


Figure 5-5: The pattern between download and citation of articles

- **Patterns of Conference Papers**

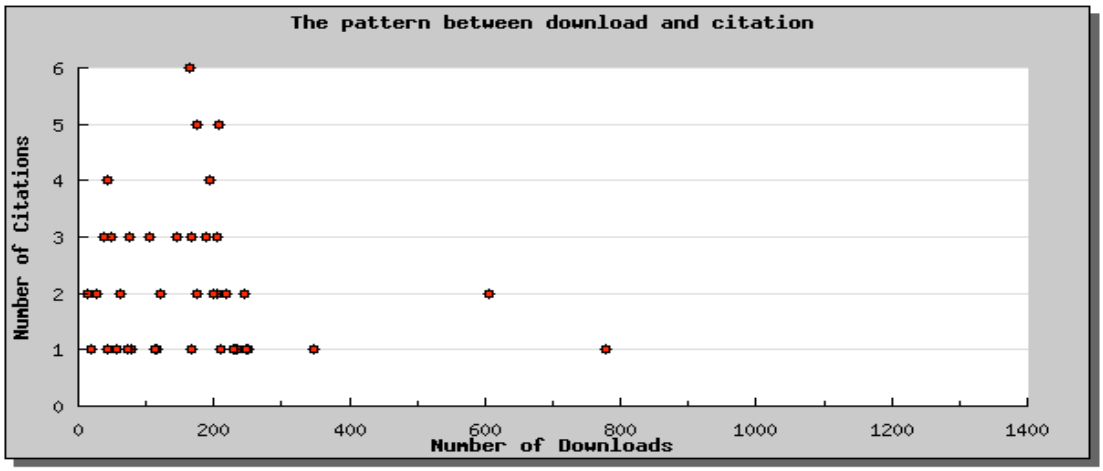


Figure 5-6: The pattern between download and citation of conference papers

- **Patterns of Books**

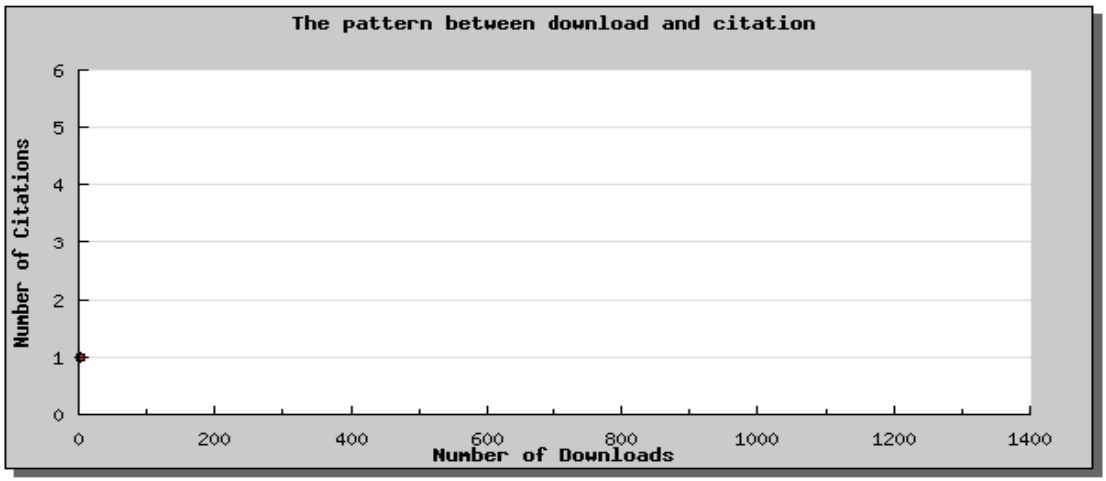


Figure 5-7: The pattern between download and citation of books

6. Conclusion

The purpose of this research is to test whether there is an identifiable causal relationship between downloads of articles and citation of the same articles from an open access repository, including the investigation of time-varying behaviours of downloading and the subsequent citations if time permits.

The software system was developed by bringing together of both downloading data and citation data for documents in UTas repository. The results derived from the data produced by the software system developed suggest that there were no significant patterns between downloads and citations for *all* the documents in the UTas repository, while there was a general trend for higher citations to be associated with higher download frequency if only the theses deposited in the repository were taken into account. The general distribution of downloads and citations are relatively scattered.

In addition, with limitation of the length of honours program, the time-varying behaviours of downloads and citations, which would require the data to be accumulated over periods of years, cannot be examined. However, the research conducted in this thesis can be used as a basis for further research as described in the following chapter.

7. Further work

The software system has been successfully implemented to test the initial hypothesis, even though not all of the aspects of them have been examined due to the time limitation issue. Therefore, many areas with potential for further research are suggested to be conducted in the future.

7.1 Time-varying behaviour

Obviously, the success of investigating the time-varying behaviour of downloads from open access repositories, and the subsequent citation would require a greater amount time, at least a couple of years according to the similar extant research.

If time permits, the downloading data and citation data will be tracked at a longer regular interval rather than only once a week along a long term, to investigate more significant changes from the data. The analysis, then, will be conducted into the download latency, which refers to the time between an article being deposited and the later downloaded, as well as the citation latency for the same article, from which the correlation between the time-varying behaviour of downloads of articles in an open access repositories, and similar characteristics of citations of the same articles will be investigated. The later citation impact therefore may be possible to predict from the known downloads, as least on average.

7.2 Aggregated Data Analysis

It would also be entirely worthwhile to cover more monitoring repositories, rather than only UTas eprints itself, to reflect on the general trend of time-varying behaviour of downloads and citations for articles in open access repositories, which can be implemented by simply providing more or other URLs of university repositories to the software system to monitor and perhaps some minor changes to the codes depending on the HTML structures of the target monitoring repositories.

In addition to this, the data could be collected and aggregated by the disciplines to which articles belong and the journals in which they are published. The analysis will undertake to test whether the previous identified patterns between downloads and citations will vary by different categories.

Since the article quality may also be a factor, this will also be included in the research. The quality could be measured in terms of the Journal Citation Impact Factor (CIF), representing the number of times an average recent article is cited in a given year, of the journal in which it is published, or the Journal Usage Impact Factor (UIF) which represents the number of times an average recent article is downloaded in a given year (Rowlands & Nicholas 2007). Thus, the behaviours for a range of articles from high quality to low quality could also be examined.

Furthermore, it would be of interest to find out whether there is an identifiable pattern between the articles that published much earlier and those of recent articles.

7.3 Merging into UTas eprints

Following a request from a librarian of UTas eprints, the citation track feature that provided by the software system could be merged with the UTas repository, so that not only the downloading data, but also the citation data will be track on through their web pages. There are a variety of ways to present this.

7.4 Miscellaneous

The identification of whether the thesis has been self-cited could be defined, so that the behaviours of self-citation for theses could be examined. Also, the reason of why some peer-reviewed journal articles have very high downloading data would be of interest to investigate.

8. References

Advanced Scholar Search Tips, 2007, Google, viewed 3 Nov 2007, <http://scholar.google.com/intl/en/scholar/refinesearch.html>.

Atkins, H 1999, 'The ISI® Web of Science® - Links and Electronic Journals', *D-Lib Magazine*, vol. 5, no. 9, pp. 1082-9873.

Bakkalbasi, N, Bauer, K, Glover, J & Wang, L 2006, 'Three options for citation tracking: Google Scholar, Scopus and Web of Science', *Biomedical Digital Libraries*, vol. 3, no. 7, pp. 1-8.

Brody, T 2006, 'Evaluating Research Impact through Open Access to Scholarly Communication', PhD thesis, University of Southampton.

Brody, T, Harnad, S & Carr, L 2006, 'Earlier Web usage statistics as predictors of later citation impact', *Journal of the American Society for Information Science and Technology*, vol. 57, no. 8, pp. 1060 - 1072.

Chan, L, Cuplinskas, D, Eisen, M, Friend, F, Genova, Y, Guédon, J-C, Hagemann, M, Harnad, S, Johnson, R, Kupryte, R, Manna, ML, Rév, I, Segbert, M, Souza, Sd, Suber, P & Velterop, J 2002, *Budapest Open Access Initiative* Budapest Open Access Initiative, viewed 19 June 2007, <http://www.soros.org/openaccess/read.shtml>.

Charles W. Bailey, J 2005, *OPEN ACCESS BIBLIOGRAPHY*, Association of Research Libraries, Washington, D.C.

Crontab - Quick reference, Admin's Choice, viewed 30 Oct 2007, <http://www.adminschoice.com/docs/crontab.htm>.

Elsevier 2007, *Scopus Overview: What is it?*, Elsevier B.V., viewed 29 Oct 2007, <http://info.scopus.com/overview/what/>.

ePrints home, UTAS ePrints, viewed 30 Oct 2007, <http://eprints.utas.edu.au/>.

Garfield, DE 1955, 'Citation Indexes for Science', *Science*, vol. 122, no. 3159, pp. 108-111.

---- 1994, 'The Concept of Citation Indexing', viewed 14 Oct 2007, <http://scientific.thomson.com/free/essays/citationindexing/concept/>.

Google Scholar Help, 2007, Google, viewed 2 Nov 2007, <http://scholar.google.com/intl/en/scholar/help.html>.

Group, PD 2007, *PHP Manual*, The PHP Group, viewed 11 June 2007, <http://www.php.net/manual/en/index.php>.

GuinuX 2002, *Class: Advanced HTTP Client*, PHP Developer's Network, viewed 21 Oct 2007, <http://phpclasses.sitehost.co.nz/browse/package/576.html>.

Hajjem, C, Harnad, S & Gingras, Y 2005, 'Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact', *IEEE Data Engineering Bulletin*, vol. 28, no. 4, pp. 39-47.

Harnad, S & Brody, T 2004, 'Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals', *D-Lib Magazine*, vol. 10, no. 6, pp. 1082-9873.

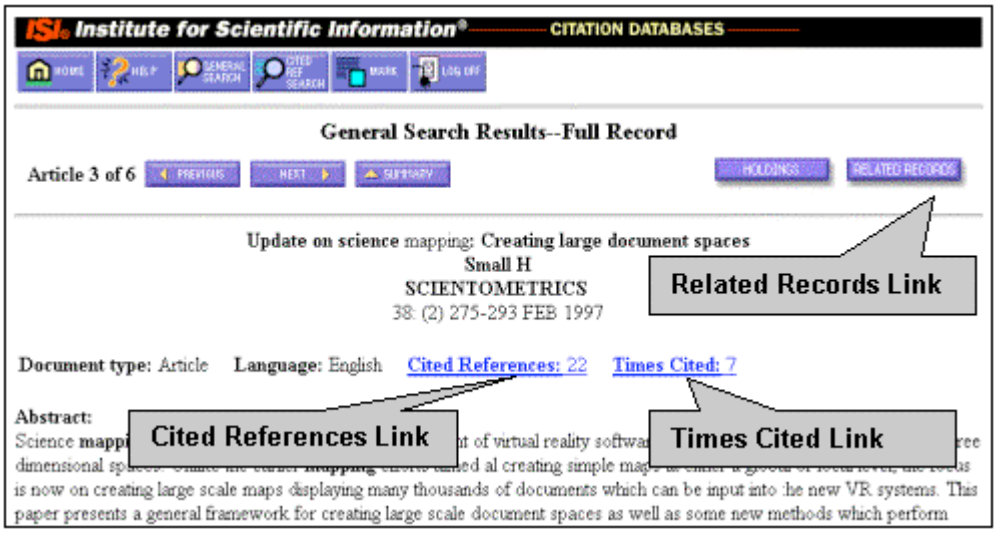
Harnad, S, Brody, T, Vallières, F, Carr, L, Hitchcock, S, Gingras, Y, Oppenheim, C, Stamerjohanns, H & Hilf, ER 2004, 'The Access/Impact Problem and the Green and Gold Roads to Open Accessstar, open', *Serials Review* vol. 30, no. 4, pp. 310-314.

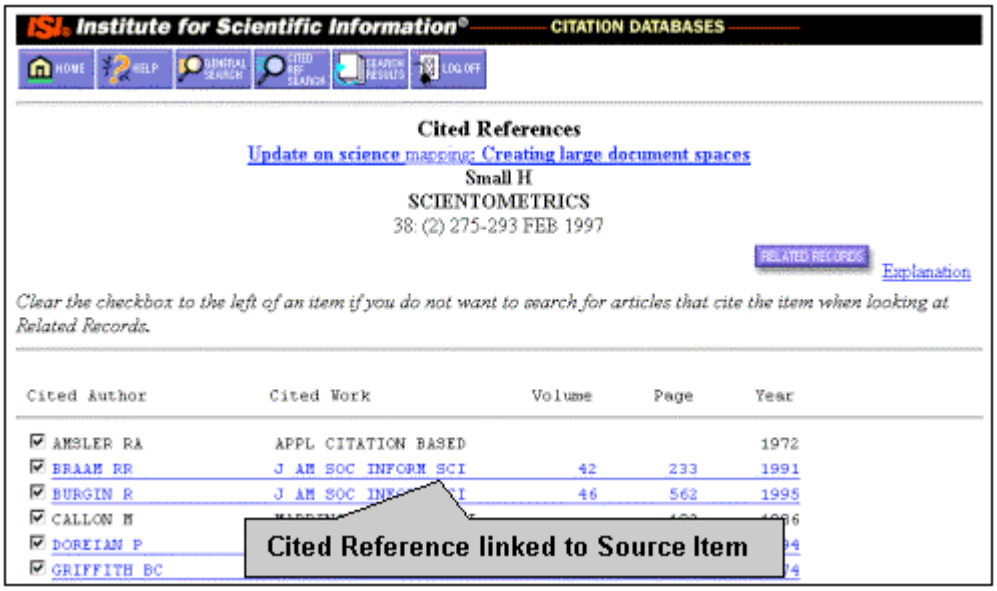
HTML Parser, 2006, SOURCEFORGE.NET, viewed 17 Oct 2007, <http://htmlparser.sourceforge.net/>.

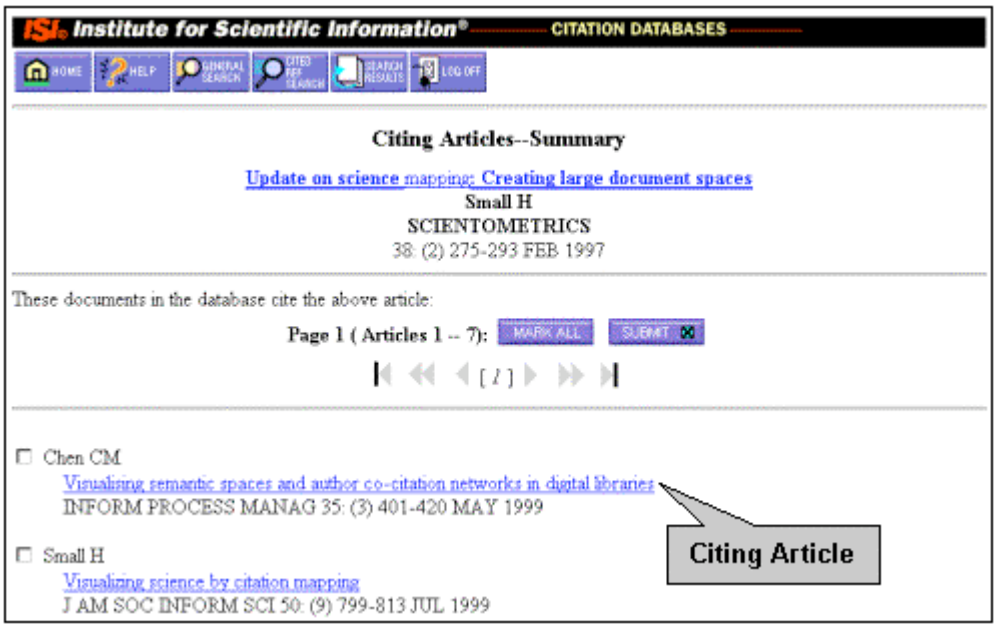
I.Meho, L & Yang, K 2007, 'Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science Versus Scopus and Google Scholar',

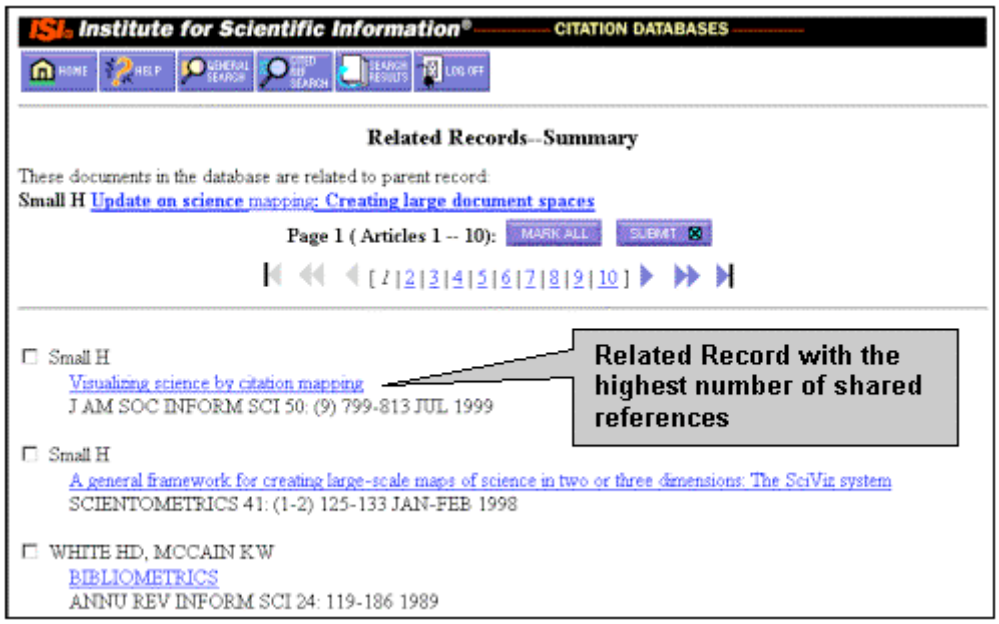
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, vol. 58, no. 13, pp. 1-21.
Jakarta Commons HttpClient, 2007, Apache Software Foundation, viewed 10 June 2007, <http://jakarta.apache.org/httpcomponents/httpclient-3.x/>.
John, J 2006, *htmlSQL*, viewed 21 Oct 2007, <http://www.jonasjohn.de/lab/htmlsql.htm>.
Library, HHS 2006, *Scopus*, The George Washington University, viewed 30 Oct 2007, <http://www.gwumc.edu/library/tutorials/PDF/Scopus_factsheet.pdf>.
Noruzi, A 2005, 'Google Scholar: The New Generation of Citation Indexes', *LIBRI - COPENHAGEN-*, vol. 55, no. 4, pp. 170-180.
OAI for Beginners: Overview, 2003, Open Archives Initiative, viewed 19 June 2007, <http://www.oaforum.org/tutorial/english/page1.htm>.
Poynder, R 2004, 'Ten Years After', *Information Today*, vol. 21, no. 9, p. 1.
Quint, B 2004, *Google Scholar Focuses on Research-Quality Content* Information Today, Inc., viewed 30 Oct 2007, <http://newsbreaks.infotoday.com/nbreader.asp?ArticleID=16324>.
Raynard, M 2007, *Scopus: What is it?*, University of Manitoba, viewed 30 Oct 2007, <http://myuminfo.umanitoba.ca/index.asp?sec=1008&too=100&dat=10/18/2007&sta=3&wee=3&eve=8&npa=13519>.
Rowlands, I & Nicholas, D 2007, 'The missing link: journal usage metrics', *Aslib Proceedings*, vol. 59, no. 3, pp. 222-228.
Sale, A 2005, *Open Access*, AuseAccess viewed 7 July 2007, <http://leven.comp.utas.edu.au/AuseAccess/pmwiki.php?n=General.OpenAccess>.
---- 2007, *Activity: Australian Repositories*, viewed 30 Oct 2007, <http://leven.comp.utas.edu.au/AuseAccess/pmwiki.php?n=Activity.AustralianRepositories>.
Self-Archiving FAQ, 2007, EPrints for Digital Repositories, viewed 30 Oct 2007, <http://www.eprints.org/openaccess/self-faq/#publisher-forbids>.
Testa, J & McVeigh, ME 2004, *The impact of open access journals: A citation study from Thomson ISI*, The Thomson Corporation, viewed 29 Oct 2007, <http://www.thomsonscientific.com/media/presentrep/acropdf/impact-oa-journals.pdf>.
Weibel, S 1999, 'The Dublin Core: A Simple Content Description Model for Electronic Resources', *NFAIS Newsletter*, vol. 40, no. 7, pp. 117-119.
Welling, L & Thompson, L 2003, *PHP and MySQL Web Development*, Second edn, Sams Publishing, Indianapolis.
What is JpGraph?, 2007, Aditus Consulting, viewed 30 Oct 2007, <http://www.aditus.nu/jpgraph/index.php>.
Willison, S 2003, *HttpClient - a PHP Web Client Class* Incutio Limited, viewed 15 Oct 2007, <http://scripts.incutio.com/httpclient/>.
Yancey, R 2005, *FIFTY YEARS OF CITATION INDEXING AND ANALYSIS*, THOMSON, viewed 10 Oct 2007, <http://scientific.thomson.com/news/newsletter/2005-08/8289803/>.

Appendix A – How ISI Web of Science Works






Institute for Scientific Information
CITATION DATABASES
HOME
HELP
GENERAL SEARCH
CITED REF SEARCH
SEARCH RESULTS
LOG OFF
Citing Articles--Summary
Update on science mapping: Creating large document spaces
Small H
SCIENTOMETRICS
38: (2) 275-293 FEB 1997
These documents in the database cite the above article:
Page 1 (Articles 1 -- 7):
MARK ALL
SUBMIT
[1]
Chen CM
Visualising semantic spaces and author co-citation networks in digital libraries
INFORM PROCESS MANAG 35: (3) 401-420 MAY 1999
Small H
Visualizing science by citation mapping
J AM SOC INFORM SCI 50: (9) 799-813 JUL 1999
Citing Article


Institute for Scientific Information
CITATION DATABASES
HOME
HELP
GENERAL SEARCH
CITED REF SEARCH
SEARCH RESULTS
LOG OFF
Related Records--Summary
These documents in the database are related to parent record:
Small H Update on science mapping: Creating large document spaces
Page 1 (Articles 1 -- 10):
MARK ALL
SUBMIT
[1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10]
Small H
Visualizing science by citation mapping
J AM SOC INFORM SCI 50: (9) 799-813 JUL 1999
Related Record with the highest number of shared references
Small H
A general framework for creating large-scale maps of science in two or three dimensions: The SciViz system
SCIENTOMETRICS 41: (1-2) 125-133 JAN-FEB 1998
WHITE HD, MCCAIN KW
BIBLIOMETRICS
ANNU REV INFORM SCI 24: 119-186 1989

Appendix B – CD-ROM

The following items are available on the accompanying CD-ROM:

- Source Code
  - Data Collection
    1. CHttp.php – HTTPClient
    2. CHtmlSql.php – HTML Parser
    3. httptester.php – Implementation file
    4. dbfuncs.inc – Database Connection
  - Data Presentation
    1. search.php – showing the records from database
    2. history.php – showing all the history records for each document
    3. statistic.php – showing some statistic graphs
- Monitoring Data in Excel (Note: Document type data is valid since 4 Nov 07)
- SQL Database